Exploratory Data Analysis

LIMITED PRELIMINARY EDITION

Volume I

by JOHN W. TUKEY

ADDISON-WESLEY PUBLISHING COMPANY
READING, MASSACHUSETTS
MENLO PARK, CALIFORNIA · LONDON · DON MILLS, ONTARIO

010.00

EXPLORATORY DATA ANALYSIS

VOLUME I

Limited Preliminary Edition

by

John W. Tukey

Princeton University
and
Bell Telephone Laboratories, Incorporated

The writing of this work together with some of the research reported herein, was supported in part by the United States Government. Reproduction, translation, publication, use, and disposal in whole or in part, by or for the United States Government is permitted.

Copyright © 1970 by Addison-Wesley Publishing Company, Inc. Philippines copyright 1970 by Addison-Wesley Publishing Company, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher. Printed in the United States of America. Published simultaneously in Canada.

TABLE OF CONTENTS

Volume 1, Preliminary Edition

(For details, see tables of contents of individual chapters.)

Preface

Introduction to the preliminary edition
About the problems

1. Scratching-down numbers (stem-and-leaf displays)

2. Easy summaries (ranks and depths)

3. Easy re-expression (quick logs, roots and reciprocals)

4. Splitting-up numbers (residuals)

5. Visual summaries (schematic plots)

6. Skipping analyses (u-, t-, and s-values)

7. Effective comparison (well-chosen expression)

7a. Means and midmeans with less arithmetic

8. Plots of relationship (straightening and levelling)

9. Straightening out plots (using three points)

10. Tilts are there to be taken away (arithmetic)

11. Which line? (y vs x, x vs y, compromise and anchored lines)

12. Looking at batches of (x,y) points (2-way schematic plots)

13. Too many points? (cutting down large (x,y) batches)

14. Bends are also there to be taken away.

15. Fitting, insulation, transparency (matters of principle)

PREFACE

This book is based on a deep conviction: it is well to understand what you can do before you learn how to measure how well you seem to have done it. The principles and procedures of what we may call confirmatory data analysis are one of the great intellectual products of this century. In their simplest form, these principles and procedures look at a sample -- and at what that sample has told us about the population from which it came -- and assess the precision with which our inference from sample to population is made. We can no longer get along without confirmatory data analysis. But we need not start with it.

The best way to understand what can be done is no longer -- if it ever were -- to ask what things could, in the present state of our skill and techniques, be confirmed. Yet many recent books have accepted not only this straight-jacket but another still more binding: the notion that what we do in dealing with data should be "best possible" according to some very restrictive assumptions, assumptions which we know we cannot check.

Once upon a time, statisticians only explored. They they learned to confirm exactly -- a few things under a few circumstances. Their techniques became, inevitably, less flexible. The connection of their techniques with past insights was weakened and anything to which a confirmatory procedure was not explicitly attached was denigrated by the word "descriptive".

We no longer need take this view. We can be flexible in our explorations, and still -- as the companion account on Confirmatory Data Analysis, yet to be written, will show -- we can apply confirmation procedures to the results of our explorations.

Exploratory data analysis is a first step. Both steps are important. Neither ought to constrain the other. Neither need do so.

It is a pleasure to acknowledge support, guidance, cooperation, and hard work. The Army Research Office (Durham), through a contract with Princeton University, and the Bell Telephone Laboratories have supported the writing financially. Charles P. Winsor taught the writer, during the 1940's, many things about data analysis that were not in the books. Although its formal beginning came after his death, it owes much to S. S. Wilks, whose leadership for statistics in Princeton made possible the gaining of the insights on which it is based. Careful reading, in various versions, by Dr. Leonard Sternberg and Mr. David Hoaglin have removed many errors, unclarities and awkwardnesses. (As always, the writer is solely responsible for those that remain.) Careful and skilled typing, mainly by Miss Elizabeth L. LaJeunesse, later Mrs. Elizabeth Dutka, and more recently by Mrs. Mary Bittrich has been an essential.

It is important to read the Introduction to the Preliminary Edition which follows.

JOHN W. TUKEY

Princeton, New Jersey
July 6, 1970

INTRODUCTION TO THE PRELIMINARY EDITION

This is a slightly revised form of the third version of this account. Each version has been used in class in successive years. Each time enough has been learned to call for extensive revision and rearrangement. This was as true of the third trial as of the other two. This time, however, the signals are for a substantial reorientation rather than for the insertion of new material or the discussion of new techniques.

The preface has hinted at the separation of exploration from confirmation. What has not yet been recognized in this preliminary edition is the separation of breadth of possibility from seriousness of analysis. The First Edition, as seen dimly today, will try to bring the breadth of possibility into its first volume, making clear at the beginning some of the diversity of ways in which data can be explored. The second volume can then go back over these approaches in turn, explaining how each can be made more carefully -- made in a way suited to controversy or publication.

We do have three obligations to our data, obligations we ought to meet one after another. We need to explore the data flexibly -- here flexibility can be enhanced by simplicity without exhaustiveness. We need to carry out the chosen analyses carefully and exhaustively -- especially if we are to

use them in a controversy. (All publication is potentially part of a controversy, just as all use as evidence is potentially part of an adversary position.) We need to assess their quality -- here is the place for Confirmatory Data Analysis.

The present version mixes exploration and care together, treating each major topic from both points of view before passing on to the next. The teacher may wish to move closer to the structure planned for the First Edition by going quite selectively through both volumes of the preliminary edition. He cannot do this as well as he might; this version was not planned to let him do this -- but he can try. The guide to short-cutting that follows offers some help.

ABOUT THE PROBLEMS

The teacher needs to be careful about assigning problems. Not too many, please. They are likely to be longer than you think. The number supplied at the ends of the chapters is to accommodate diversity of interest, not to keep everyone busy.

Besides the length of our problems, both teacher and student need to realize that many problems do not have a single "right answer". There can be many ways to approach a body of data. Not all are equally good. For some this may be clear, but for others we may not be able to tell from

this body of data which is preferred. Even several bodies of data about very similar situations may not be enough to show which should be preferred. Accordingly, it will often be quite reasonable for different analysts to reach somewhat different analyses.

Yet more -- to unlock the analysis of a body of data, to find the good way or ways to approach it, may require a key, whose finding is a creative act. Not everyone can be expected to create the key to any one situation. And, to continue to paraphrase Barnum, no one can be expected to create a key to each situation he or she meets.

An eminent statistician once visited an eminent prep school teacher of physics. After the laboratory group was well into the experiment, the statistician hinted that it might be interesting to collect all the students results and display them together. "Oh", said the teacher, "but some of them will have gotten the wrong answer." The usually free-spoken statistician, whom long experience had taught to expect variation in everything, dared not trust himself to reply. To learn about data analysis it is right that each of us try many things that do not work -- that we tackle more problems than we make expert analyses of. We often learn less from an expertly done analysis than from one where we missed an opportunity -- at least until we were told about it. Each teacher needs to recognize this in grading and commenting on problems.

GUIDELINES FOR SHORTCUTTING AND DELAYING

Chapter	Various levels of shortcutting									
	Extreme		Intermediate		Moderate		Mild		None	
	Sections	*	Sections	*	Sections	*	Sections	*	Sections	*
1	A to E	13	A to G	20	A to G	20	A to G	20	A to G	20
2	A to E	12	A to E	12	A to E	12	A to E,G	14	A to G	16
3	A to E	18	A to E	18	A to E	18	A to E	18	A to G	22
4	A to D	15	A to D	15	A to D	15	A to F	22	A to F	22
5	------	--	A to D	17	A to D	17	A to G	25	A to G	25
6	------	--	------	--	------	--	A to G	15	A to G	15
7	------	--	------	--	A to B	8	A to B	8	A to E	13
7z	------	--	------	--	A to B	5	A to B	5	A to B	5
8	A to E	24	A to E	24	A to E	24	A to E	24	A to E	24
9	A to G	24	A to G	24	A to G	24	A to G	24	A to G	24
10	**	10	A to C	19	A to D,F	24	A to F	26	A to G	28
11	A to C	8	A to C	8	A to C	8	A to C	8	A to F	22
12	------	--	------	--	A to F	22	A to F	22	A to F	22
13	------	--	A to E	7	A to E	7	A to E,G	10	A to G	16
14	------	--	------	--	------	--	A to D	15	A to D	15
15	------	--	------	--	------	--	------	--	A to C	13
		124		164		204		256		302

* Pages of text (not exhibits), Sections H and P, which would be covered, omitted from count.

** Pages 1-7, 21-23, 26-27.

The extreme short-cut will ordinarily only be taken in extreme haste to reach the two-way analyses of preliminary volume 2.

If any of the more vigorous short-cuts are taken, it will be natural to return to chapter 6 after covering parts of preliminary volume 2.